Estatística Inferencial

Prof. Wagner Hugo Bonat

Departamento de Estatística
Universidade Federal do Paraná

DEST
Departamento de Estatística

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

Verossimilhança e suas derivadas

Notação e definições (relembrando)

- $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$: v.a.'s independentes e idênticamente distribuídas.
- $Y_i \sim f(\boldsymbol{\theta})$ onde f denota a função densidade de probabilidade ou função de probabilidade e $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\top$ é um vetor de p parâmetros populacionais.
- $\mathbf{y} = (y_1, \ldots, y_n)^\top$ denota o vetor de valores observados da v.a. Y.
- **Estatística** - Uma estatística é uma variável aleatória $\mathrm{T} = \mathrm{t}(\mathbf{Y})$, onde a função $\mathrm{t}(\cdot)$ não depende de $\boldsymbol{\theta}$.
- **Estimador** - Uma estatística T é um estimador para θ se o valor realizado $\mathrm{t} = \mathrm{t}(\boldsymbol{y})$ é usado como uma estimativa para o valor de θ.
- **Distribuição amostral** - A distribuição de probabilidade de T é chamada de distribuição amostral do estimador $\mathrm{t}(\mathbf{Y})$.

Função de verossimilhança (caso uniparamétrico)

- **Função de verossimilhança** - Seja $\mathbf{y}$ um vetor $n \times 1$ representando uma realização de um vetor aleatório Y com função de probabilidade ou densidade probabilidade $\mathrm{f}(\mathrm{Y}, \boldsymbol{\theta})$, onde θ denota um parâmetro, com $\theta \in \Theta$, sendo Θ o respectivo espaço paramétrico. A função de verossimilhança ou simplesmente verossimilhança para θ dado os valores observados $\mathbf{y}$ é a função aleatória $\mathrm{L}(\theta|\mathbf{y}) \equiv \mathrm{f}(\mathrm{Y}, \theta)$.
- Caso discreto:

$$\mathrm{L}(\theta|\mathbf{y}) \equiv \mathrm{P}_\theta(\mathrm{Y} = \mathbf{y}).$$

- Caso contínuo (simplificado)

$$\mathrm{L}(\theta|\mathbf{y}) \approx \prod_{i=1}^{n} \mathrm{f}(y_i, \theta).$$

Verossimilhança - Condições de regularidade

- O parâmetro θ é **identificável**. Isso significa que se $f(\theta_1|\boldsymbol{y}) = f(\theta_2|\boldsymbol{y})$ para quase todos $y \in \Re$, então $\theta_1 = \theta_2$.
- Para quase todo significa que a condição não é verdadeira para um conjunto de $\boldsymbol{y}$ com probabilidade zero de ocorrência.
- O suporte de $f(\theta|\boldsymbol{y})$ é o mesmo para todo $\theta \in \Re$.
- O verdadeiro valor do parâmetro θ_0 pertence ao interior de Θ.
- $f(\theta|\boldsymbol{y})$ é duas vezes continuamente diferenciável com relação θ para quase todo $\boldsymbol{y} \in \Re$.
- $\frac{\partial}{\partial \theta}$ e $\int$ (caso contínuo) ou $\frac{\partial}{\partial \theta}$ e $\sum$ (caso discreto) podem ser intercambiada.

Log-Verossimilhança (caso uniparamétrico)

- A função de log-verossimilhança é a função estocástica $l(\theta) : \Theta \rightarrow \Re$ definida por

$$l(\theta|\boldsymbol{y}) = \log\left(L(\theta|\boldsymbol{y})\right).$$

- No caso iid, tem-se

$$l(\theta|\boldsymbol{y}) = \sum_{i=1}^{n} \log\left(L(\theta|y_i)\right).$$

- $l(\theta|\boldsymbol{y}) = -\infty$ quando $L(\theta) = 0$, mas isso ocorre quando $f(y_1, \ldots, y_n|\theta) = 0$ que tem probabilidade de ocorrência igual a zero.

Desigualdade de Jensen

- **Desigualdade de Jensen**: Seja g uma função estritamente convexa e Y uma v.a com $\mathrm{E}(|Y|) < \infty$ tal que a distribuição de Y é não degenerada. Então,
  - $g(\mathrm{E}(Y)) < \mathrm{E}(g(Y))$. Por outro lado, se g é estritamente côncava, então
  - $g(\mathrm{E}(Y)) > \mathrm{E}(g(Y))$.
- Lembrete:
  - Função convexa $f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$.
  - Função côncava $f(tx + (1-t)y) \geq tf(x) + (1-t)f(y)$.

Desigualdade de Jensen e Máxima Verossimilhança

- Teorema: Seja θ_0 o verdadeiro valor do parâmetro. Então,

$$P_{\theta_0}(\mathrm{L}(\theta_0|\boldsymbol{y}) > \mathrm{L}(\theta|\boldsymbol{y})) \to 1, \quad \text{quando} \quad n \to \infty.$$

- Interpretação: $\mathrm{L}(\theta_0) > \mathrm{L}(\theta)$ com alta probabilidade para n grande.
- Assim, $\mathrm{L}(\theta)$ vai tender a ter o seu máximo próximo a θ_0, o verdadeiro valor de θ.
- Motiva a ideia de estimação por máxima verossimilhança.
- Demonstração (vídeo separado opcional).

Função escore e Informação de Fisher (caso uniparamétrico)

- Função escore para θ (efficient score)

$$\begin{aligned} \mathrm{U}(\theta|\boldsymbol{Y}) &= \mathrm{U}(\theta|Y_1, \ldots, Y_n) \\ &= \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \mathrm{l}(\theta, Y_i). \\ &= \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f(\theta, Y_i). \end{aligned}$$

- Informação de Fisher ou Informação esperada

$$\mathrm{I_E}(\theta) = \mathrm{Var}(\mathrm{U}(\theta|\boldsymbol{Y})).$$

- Informação de Fisher também é chamada de *intrinsic accuracy*.

Igualdades de Bartlett (caso uniparamétrico)

- Sob condições de regularidade, tem-se
  - Primeira igualdade: $\mathrm{E}(\mathrm{U}(\theta|\boldsymbol{Y})) = 0$.
  - Segunda igualdade: $\mathrm{I}_{\mathrm{E}}(\theta) = -\mathrm{E}(\mathrm{l}''(\theta|\boldsymbol{Y})) = -\mathrm{E}(\mathrm{U}'(\theta|\boldsymbol{Y}))$.
- Implicação: $\mathrm{Var}(\mathrm{U}(\theta|\boldsymbol{Y})) = \mathrm{E}(\mathrm{U}(\theta|\boldsymbol{Y})^2)$.
- Demonstração.
- Exercício: Sejam $Y_i \sim P(\lambda)$ iid, para $i = 1, \ldots, n$. Verifique as igualdades de Bartlett.

Informação observada

- Informação observada para θ

$$\mathrm{I}_{\mathrm{O}}(\theta) = -\mathrm{l}''(\theta|\boldsymbol{Y}).$$

  - Note que $\mathrm{I}_{\mathrm{E}}(\theta) = \mathrm{E}(\mathrm{I}_{\mathrm{O}}(\theta))$.
  - Além disso, pela lei dos grandes números

$$\mathrm{I}_{\mathrm{O}}(\theta) \xrightarrow{P} \mathrm{I}_{\mathrm{E}}(\theta) \quad \text{quando} \quad n \rightarrow \infty.$$

- Exercício: Sejam $Y_i \sim B(p)$ iid, para $i = 1, \ldots, n$. Encontre a informação observada e esperada e mostre que a informação esperada coincide com a variância da função escore.

Desigualdade de Cramér-Rao (caso uniparamétrico)

- Teorema: Se $T(Y_1, \ldots, Y_n)$ é um estimador não viciado para θ, então

$$\mathrm{Var}(T) \geq \mathrm{I}_{\mathrm{E}}(\theta)^{-1}.$$

- A quantidade $\mathrm{I}_{\mathrm{E}}(\theta)^{-1}$ é chamado de limite inferior de Cramér-Rao.
- Um estimador não viciado é chamado eficiente se $\mathrm{Var}(T) = \mathrm{I}_{\mathrm{E}}(\theta)^{-1}$.
- Demonstração.

Vetor de parâmetros

Função de verossimilhança (caso multiparamétrico)

- Função de verossimilhança $\mathrm{L} : \Theta \to [0, \infty]$ é uma função aleatória de um **vetor** de argumentos definida por

$$\mathrm{L}(\boldsymbol{\theta}|\boldsymbol{y}) = \prod_{i=1}^{n} f(y_i, \boldsymbol{\theta}), \quad \text{para} \quad \boldsymbol{\theta} \in \Theta.$$

- Função de log-verossimilhança $\mathrm{l} : \Theta \to \Re$ é uma função aleatória de um **vetor** de argumentos definida por

$$\mathrm{l}(\boldsymbol{\theta}|\boldsymbol{y}) = \sum_{i=1}^{n} \log f(y_i, \boldsymbol{\theta}).$$

- Equivalentemente,

$$\mathrm{l}(\boldsymbol{\theta}|\boldsymbol{y}) = \log \mathrm{L}(\boldsymbol{\theta}|\boldsymbol{y}).$$

Vetor escore

- O vetor escore $\mathrm{U}(\boldsymbol{\theta}|\boldsymbol{y}) : \Theta \to \Re^p$ é um vetor aleatório $p \times 1$ definido por

$$\mathrm{U}(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{\partial \mathrm{l}(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \boldsymbol{\theta}} = \begin{pmatrix} \frac{\partial \mathrm{l}(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \theta_1} \\ \vdots \\ \frac{\partial \mathrm{l}(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \theta_p} \end{pmatrix}.$$

- Notação popular em termos de gradiente

$$\mathrm{U}(\boldsymbol{\theta}|\boldsymbol{y}) = \nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}|\boldsymbol{y}).$$

- Notação j-ésimo componente de $\mathrm{U}(\boldsymbol{\theta}|\boldsymbol{y})$ por $\mathrm{U}_j(\boldsymbol{\theta}|\boldsymbol{y})$.

Esperança do vetor escore

- O vetor escore satisfaz as igualdades de Bartlett, ou seja,

$$\mathrm{E}(\mathrm{U}(\theta|Y)) = \mathbf{0},$$

isso significa que

$$\mathrm{E}(\mathrm{U}_j(\theta|Y)) = 0, \quad \text{para} \quad j = 1, \ldots, p.$$

Matriz de informação esperada

- A matriz $p \times p$ definida por

$$
\begin{aligned}
\mathrm{I}(\boldsymbol{\theta}) &= \mathrm{Var}(\mathrm{U}(\boldsymbol{\theta}|\boldsymbol{Y})) \\
&= \mathrm{E}(\mathrm{U}(\boldsymbol{\theta}|\boldsymbol{Y})\mathrm{U}^\top(\boldsymbol{\theta}|\boldsymbol{Y})).
\end{aligned}
$$

  é chamada de matriz de informação esperada.
- As entradas j e k são expressadas por

$$
\begin{aligned}
\mathrm{I}_{jk}(\boldsymbol{\theta}) &= \mathrm{Cov}(\mathrm{U}_j(\boldsymbol{\theta}|\boldsymbol{Y}), \mathrm{U}_k(\boldsymbol{\theta}|\boldsymbol{Y})) && (1) \\
&= \mathrm{E}(\mathrm{U}_j(\boldsymbol{\theta}|\boldsymbol{Y}), \mathrm{U}_k(\boldsymbol{\theta}|\boldsymbol{Y}))). && (2)
\end{aligned}
$$

Matriz de informação observada

- A matriz $p \times p$ definida por
$$\mathrm{J}(\boldsymbol{\theta}) = -\frac{\partial^2 \mathrm{l}(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}.$$
- As entradas j e k da matriz de informação observada é dada por
$$\mathrm{J}_{jk}(\boldsymbol{\theta}) = -\frac{\partial^2 \mathrm{l}(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \theta_j \partial \theta_k}.$$
- Segunda igualdade de Bartlett
$$\mathrm{I}(\boldsymbol{\theta}) = \mathrm{E}(\mathrm{J}(\boldsymbol{\theta})).$$

Desigualdade de Cramér-Rao generalizada

- Defina $\mathrm{I}^{jk}(\boldsymbol{\theta}) = \{\mathrm{I}^{-1}(\boldsymbol{\theta})\}_{jk}$. Se $\mathrm{T} = \mathrm{T}(Y_1, \ldots, Y_n)$ é um estimador não-viciado para θ_1, ou seja,
$$\mathrm{E}(\mathrm{T}) = \theta_1,$$
então
$$\mathrm{Var}(\mathrm{T}) \geq \mathrm{I}^{11}(\boldsymbol{\theta}).$$
- Demonstração análoga ao caso univariado usando a desigualdade de Cauchy-Schwarz generalizada.

Parâmetros ortogonais

- Considere um modelo estatístico parametrizado por $\boldsymbol{\theta} = (\theta_1, \theta_2)^\top$. No caso da matriz de informação de Fisher ser diagonal

$$\mathrm{I}(\boldsymbol{\theta}) = \begin{pmatrix} \mathrm{I}_{11}(\boldsymbol{\theta}) & 0 \\ 0 & \mathrm{I}_{22}(\boldsymbol{\theta}) \end{pmatrix},$$

  os parâmetros θ_1 e θ_2 são ditos **ortogonais**.
- O inverso da informação de Fisher é também diagonal

$$\mathrm{I}^{-1}(\boldsymbol{\theta}) = \begin{pmatrix} 1/\mathrm{I}_{11}(\boldsymbol{\theta}) & 0 \\ 0 & 1/\mathrm{I}_{22}(\boldsymbol{\theta}) \end{pmatrix}.$$

- Assim, $\hat{\theta}_1$ e $\hat{\theta}_2$ são assintóticamente independentes. com distribuição

$$\hat{\theta}_j \overset{a}{\sim} \mathrm{N}(\theta_j, 1/\mathrm{I}_{jj}(\boldsymbol{\theta})).$$

Parâmetros ortogonais: Generalização

- Considere um modelo estatístico parametrizado por $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)^\top$. No caso da matriz de informação de Fisher ser bloco diagonal

$$\mathrm{I}(\boldsymbol{\theta}) = \begin{pmatrix} \mathrm{I}_1(\boldsymbol{\theta}) & 0 \\ 0 & \mathrm{I}_2(\boldsymbol{\theta}) \end{pmatrix},$$

  os vetores de parâmetros $\boldsymbol{\theta}_1$ e $\boldsymbol{\theta}_2$ são ditos **ortogonais**.
- A distribuição assintótica de $\boldsymbol{\theta}_1$ é a mesma se $\boldsymbol{\theta}_2$ é considerado conhecido ou desconhecido.
- Definição similar pode ser feita usando a matriz de informação observada.

Exercício

- Sejam $Y_i \sim N(\mu, \sigma^2)$ iid para $i = 1, \ldots, n$.
  - Escreva a função de verossimilhança e log-verossimilhança.
  - Obtenha a função escore.
  - Obtenha a matriz de informação observada e esperada.
  - Obtenha a variância assintótica de $\hat{\mu}$ e $\hat{\sigma}^2$.
  - Os parâmetros μ e σ^2 são ortogonais? Justifique.